Vietnam Journal of Science and Technology ***56*** *(2C) (2018) 104-110*

APPLICATION OF MACHINE LEARNING TO FILL IN THE MISSING MONITORING DATA OF AIR QUALITY

Mac Duy Hung[1, 2, *], Nghiem Trung Dung[1], Hoang Xuan Co[3]

[1]*Hanoi University of Science and Technology, 1 Dai Co Viet road, Ha Noi, Viet Nam*

[2]*Thai Nguyen University of Technology, 3/2 road, Tich Luong ward, Thai Nguyen, Viet Nam*

[3]*VNU University of Science, 334 Nguyen Trai road, Ha Noi, Viet Nam*

[*]Email*:* *macdh@tnut.edu.vn*

Received: 10 May 2018; Accepted for publication: 21 August 2018

ABSTRACT

In this paper, three machine learning models have been applied to predict and fill in the missing monitoring data of air quality for Gia Lam and Nha Trang stations in Hanoi and Khanh Hoa respectively, including Autoregressive Moving Average (ARMA), Artificial Neural Network (ANN), and Support Vector Regression (SVR). Two air pollutants being NO_2 and PM_{10} were selected for this study. The experimental results showed that the performance of all three studied models is better than that of some traditional approaches, including Multiple Linear Regression (LR) and Spline interpolation. Besides that, ARMA, ANN and SVR can capture the fluctuation of concentrations of the selected pollutants. These results indicated that the machine learning is a feasible approach to deal with the missing of data which is one of the biggest problems of air quality monitoring stations in Viet Nam.

Keywords: air quality, ANN, ARMA, SVR, missing data.

1. INTRODUCTION

Monitoring and modeling of air quality is of ultimate significance for understanding the trend and characteristics of air pollutants. For understanding and simulating the fluctuation of an air pollutant, it is required to have the dataset of air quality which is not only long enough in time and reliable but also time-serially completion of observations. However, the continuity of time-series measurements is normally plagued with different factors including malfunction of the equipment, power cut off, not regularly maintained, etc., resulting in the gap of data points or missing data. Many statistical approaches such as linear or logistic regression, polynomial or spline interpolation/extrapolation [1, 2, 3], Kalman filter approach [4] and so on have been proposed to deal with this problem. However, none of them is effective when the number of gaps is large. In recent studies, the machine learning approaches have been successfully applied to predict values of concentrations of air pollutants [2, 5, 6, 7, 8, 9, 10, 11]. This study, therefore, aimed at the application of machine learning to fill in the gaps of air quality monitoring data in

Viet Nam focusing on autoregressive moving average (ARMA), artificial neural network (ANN) and support vector regression methods.

2. METHODOLOGY

2.1. Autoregressive moving average

Autoregressive moving average (ARMA) is a statistical model of time series analysis which combines autoregressive analysis (AR) and moving average (MA) methods. An ARMA model of x_t time series data can be defined by following equations [12].

AR component: MA component:

$$x_t = \alpha_1 x_{t-1} + ... + \alpha_p x_{t-p} + z_t \quad (1); \qquad x_t = \beta_0 z_t + \beta_1 x_{t-1} + ... + \beta_q x_{t-q} \quad (2)$$

And ARMA model:

$$x_t = \alpha_1 x_{t-1} + ... + \alpha_p x_{t-p} + z_t + \beta_1 x_{t-1} + ... + \beta_q x_{t-q} \quad (3)$$

where, α_1, …, α_p and β_1, …, β_p are the corresponding coefficients.

2.2. Artificial neural network

Artificial neural network (ANN), a mathematical model, is built based on a biological neural system that consists of three or more layers which are formed by neurons intended to simulate the learning and pattern recognition [13]. An example of typical NN structure is shown in Figure 1, which only has one hidden layer.

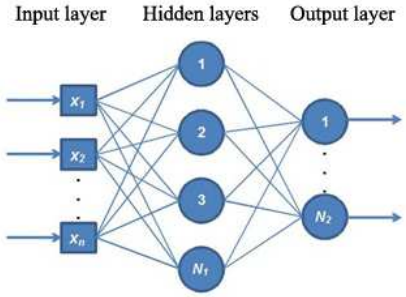


Figure 1. The structure of a typical artificial neuron network (with three layers).

The output of *i*-th neuron (x_i) is determined by the following equations (4) and (5):

$$x_i = f(\xi_i) \quad (4)$$

$$\xi_i = \sum_{j \subseteq \Gamma_i^{-1}} W_{ij}.x_j + \varepsilon_i \quad (5)$$

where, ξ_i is the potential of *i*-th neuron; $f(\xi_i)$ is called transfer function; the threshold ε_i is a weight coefficient of the connection with formally added neuron *j* (so called bias).

The equation (5) is carried out over all neuron *j* (x_j) transferring the signals to *i* (x_i) neuron.

In this study, multilayer feedforward neural network (FFNN) was used with the transfer function being sigmoidal.

2.3. Support vector regression

Support vector machine (SVM) was proposed by V. N. Vapnik [14] to deal with the problems of data classification. SVM creates a hyper plane in boundless dimensional (feature) space, which is used for classification and regression. Support vector regression (SVR) is a linear regression based on the SVM technique. A linear regression function of a given set of data x can be wrote in form $F(x) = w^T \Phi(x_i) + b$ in a feature space □, where w is the coefficient vector, b is a threshold and $\Phi(x_i)$ is a nonlinear function which maps the input x to a vector in □. In order to estimate a function F within a finite accuracy, the estimation value $\hat{F}$ of F needs to satisfy the condition $\left|F - \hat{F}\right| \le \varepsilon$, where ε is the allowably maximum deviation during the training state. The reliability of prediction values is measured by a loss function which is called ε-intensitive cost lost function [5, 15]. The SVR function for nonlinear predictions becomes the equation (6) below:

$$\hat{y} = \hat{F}(x) = \sum_{i=1}^{N} (\alpha_i - \alpha_i^*) K(x_i, x) + b \tag{6}$$

where, α and α^* are Lagrange parameters; $K(x_i, x)$ is known as a Kernel function. Any function that meets Mercer's condition [14] can be used as the Kernel function. For this study, several preliminary tests were conducted to select the Kernel function including linear, RBF and polynomial. The results indicated that the linear function is the best, therefore, it was chosen for this study.

2.4. Datasets

Data used for this study were extracted from the databases of air quality monitoring stations in Gia Lam (Hanoi, from 2014 to 2016) and Nha Trang (Khanh Hoa, from 2013 to 2015). This study focused on two primary air quality parameters, namely NO_2 and PM_{10}. Datasets for training and testing the models are extracted from these databases in periods in which data are not missing. The input data are set as the following form:

$$X = (X_{t-24}, X_{t-23}, ..., X_{t-2}, X_{t-1}, t)$$

where, X_{t-24}, …, X_{t-1} are concentrations of a studied pollutant which is needed to fill in the gaps; t is the time of gaps, i.e., to predict a missing value of a pollutant, these models need 25 values, where 24 values are concentrations of this pollutant in 24 previous hours and t is used as the variable of the concentration trend of this pollutant.

2.5. Evaluation of performance

The performance of selected models was evaluated based on statistical indicators including root mean square error (RMSE), mean absolute error (MAE) and correlation coefficient (r).

3. RESULTS AND DISCUSSIONS

In this study, testing datasets are complete segments from data sources that are assumed to be missing. The positions of these segments are set by random. In addition, the predicted value in the time t is feedback to the input of models as X_{t-1} to predict the missing concentration of next time $t+1$. The performance of gap filling of the selected models on testing datasets are presented below.

3.1. Filling in the data of NO_2

Obtained results presenting in Table 1 showed that, in almost all experiments, the performance of traditional approaches tested in this study including LR and Spline interpolation in segments with huge number of missing values is not reliable. It is because, these models try to build a function that fit the trend of the studied pollutants. However, the fluctuation of the concentrations of a pollutant in the air is a nonlinear function which is influenced by many factors (e.g., time, precursors, meteorological conditions and so on), therefore, they cannot build a fitting function for that. On the contrary, almost machine learning models including ARMA, ANN and SVR do not need to build a fitting function. Values predicting by these models are calculated from historical data in the learning process. Therefore, they can predict more accuracy. The ARMA, ANN and SVR used in this study predicted well data of Gia Lam and Nha Trang stations. The performances of these models are much better than those of LR and Spline models, not only in terms of statistical indicators but also in terms of the fluctuation trend of pollutant as presented in Figure 2. It can be seen from Figure 2 that, three models, ARMA, ANN and SVM, adapted well with the fluctuation of real NO_2 concentration. This is very important for the prediction of a parameter through time series.

Table 1. The performance of selected models on testing datasets for filling in the data of NO_2 with more than 100 missing points in Gia Lam and Nha Trang stations.

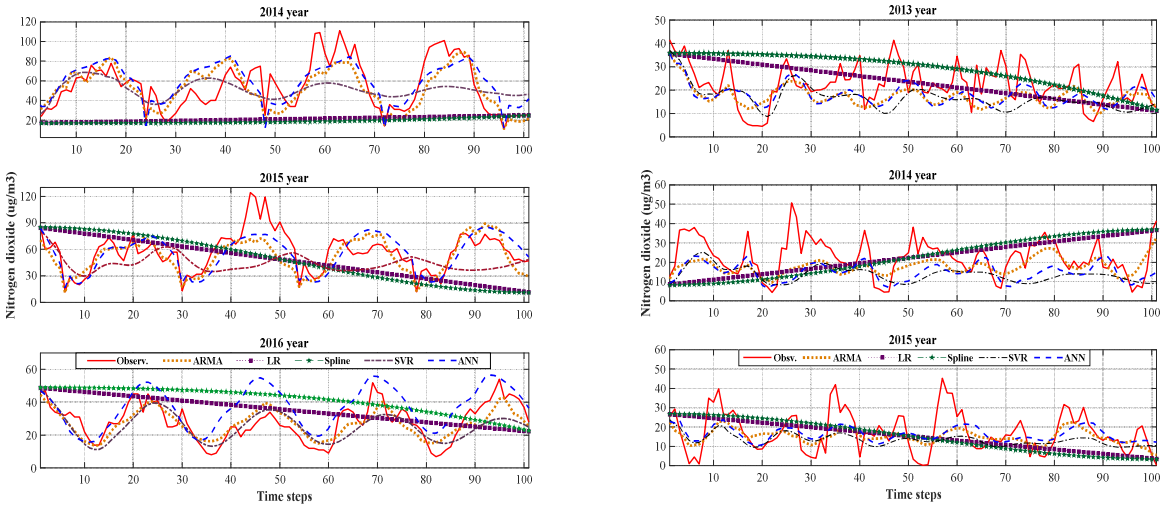
Year	Indicators	Gia Lam station (GL)					Nha Trang station (NT)				
		LR	*Spline*	*ARMA*	*ANN*	*SVR*	*LR*	*Spline*	*ARMA*	*ANN*	*SVR*
2014(GL) 2013(NT)	RMSE(μg/m^3)	41.49	43.27	16.72	17.52	22.30	11.68	12.45	8.59	8.82	9.90
	MAE(μg/m^3)	33.80	35.66	13.15	14.56	18.17	9.36	9.95	6.53	6.93	7.81
	r	0.07	-0.06	0.75	0.73	0.48	0.14	0.13	0.68	0.58	0.52
2014(GL) 2013(NT)	RMSE(μg/m^3)	32.35	35.40	16.15	15.32	26.86	13.52	14.77	9.10	11.52	13.08
	MAE(μg/m^3)	26.97	30.35	12.06	12.59	21.32	10.54	11.81	7.16	9.10	10.25
	r	-0.04	-0.03	0.74	0.74	0.30	-0.16	-0.15	0.63	0.46	0.24
2014(GL) 2013(NT)	RMSE(μg/m^3)	15.91	19.85	6.98	12.37	8.61	10.33	11.87	8.72	8.84	10.02
	MAE(μg/m^3)	13.42	17.32	5.95	10.35	7.00	8.06	9.22	7.04	7.41	7.92
	r	-0.06	-0.19	0.80	0.75	0.70	0.14	0.17	0.58	0.44	0.35

In addition, the performance of the selected models for Nha Trang station is worse than that for Gia Lam station. It is because, the number of missing data points of Nha Trang station is huge (the rates of missing points in 2013, 2014 and 2015 are 30 %, 37 % and 13 %, respectively), therefore, the models had less information to learn.

3.2. Filling in the data of PM_{10}

PM_{10} is a typical air quality parameter. It can be directly emitted into the air from local sources and/or can come from remote sources by the long-range transport. It is also formed in the air as a secondary pollutant. Besides its generation process, it can be removed from the air by the wet and dry deposition. Its level is, hence, dependent on many factors including emissions sources, meteorological conditions, topography, the concentrations of precursors such as NO_2

and SO_2, etc. The fluctuation of its concentration in the air is, therefore, very complex. In addition, the number of missing values of PM_{10} in the two stations is huge. That is why, the models cannot predict fully the fluctuation trend of PM_{10}. Thus, as can be seen from Table 2 and Figure 3, the performance of these models for PM_{10} is worse than those for NO_2 not only in terms of statistical indicators but also in terms of the ability to capture of the trend of pollutants.



(a) *(b)*

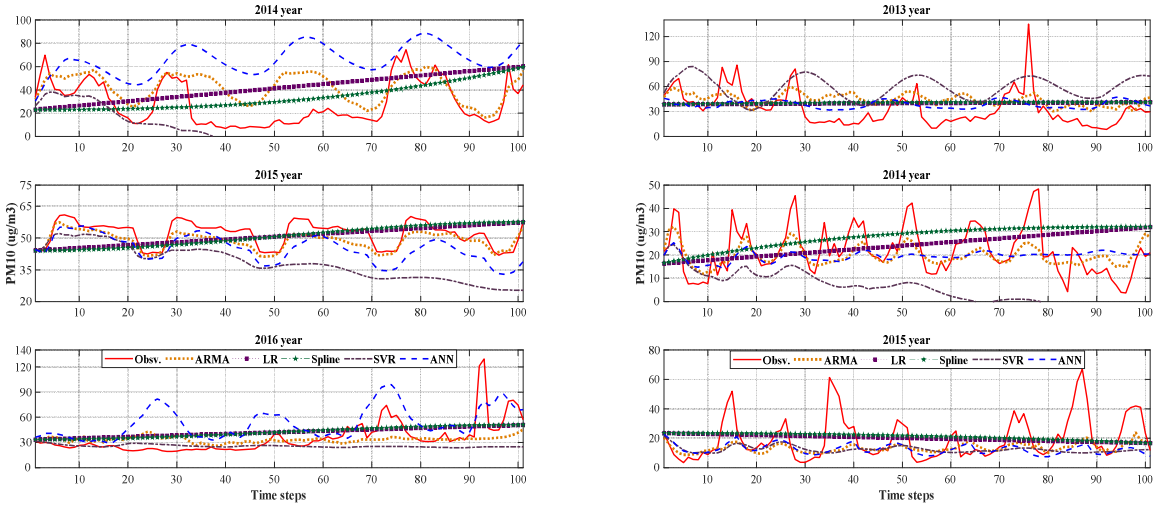
Figure 2. The comparison of measured and predicted values of NO_2 for selected models *(a)* Gia Lam station and *(b)* Nha Trang station.

Table 2. The performance of selected models on testing datasets for filling in the data of PM_{10} with more than 100 missing points in Gia Lam and Nha Trang stations.

Year	Indicators	Gia Lam station (GL)					Nha Trang station (NT)				
		LR	*Spline*	*ARMA*	*ANN*	*SVR*	*LR*	*Spline*	*ARMA*	*ANN*	*SVR*
2014(GL) 2013(NT)	RMSE(µg/m^3)	49.21	55.95	32.57	43.94	64.68	16.83	17.10	12.69	15.43	29.14
	MAE(µg/m^3)	36.84	42.30	24.43	37.63	46.65	14.25	14.43	10.11	12.88	22.96
	r	0.63	0.60	0.85	0.84	-0.58	0.59	0.57	0.68	0.57	-0.24
2014(GL) 2013(NT)	RMSE(µg/m^3)	7.50	8.03	4.35	8.14	16.05	11.88	12.63	8.01	9.83	19.95
	MAE(µg/m^3)	6.30	6.67	3.68	7.16	14.06	9.64	10.66	6.10	7.45	17.26
	r	-0.18	-0.17	0.89	0.78	0.37	-0.14	-0.02	0.69	0.35	0.26
2014(GL) 2013(NT)	RMSE(µg/m^3)	17.60	17.21	18.34	25.17	21.55	14.85	15.00	14.26	14.95	16.15
	MAE(µg/m^3)	13.98	13.98	11.56	20.39	12.61	11.99	12.30	9.32	9.52	10.44
	r	0.63	0.63	0.25	0.54	-0.31	-0.22	-0.24	0.55	0.43	-0.01

Furthermore, the results also indicated that the performance of SVR is the worst among the three models. It is contrary to what reported by previous studies [6, 7, 8] in which SVM/SVR is better than ANN and ARRMA in the prediction of air quality. This might be explained by the quality of data, the selection of inputs variable, and the different way of approach for prediction. As presented above, this study used 25 input variables (24 variables for fluctuation trend of

studied pollutants in 24 previous hours and the remaining one is used as the activity of emission source), while other forecasting studies used meteorological variables [7, 8] and precursors related to the pollutant to be predicted [6, 7, 8]. However, these results indicate that the performance of machine learning models is better than that of traditional approaches, which is consistent with the results of our previous studies [9, 10].



(a) *(b)*

Figure 3. The comparison of measured and predicted values of PM_{10} for selected models *(a)* Gia Lam station and *(b)* Nha Trang station.

4. CONCLUSIONS

Three machine learning models were used to predict the missing values of monitoring data of air quality for Gia Lam, Hanoi and Nha Trang, Khanh Hoa stations. Extensive experimental results indicated that the effectiveness of the three studied machine learning models, namely ARMA, ANN and SVR is better than that of traditional approaches such as LR and Spline interpolation. It is found that the quality of dataset in terms of missing data points significantly influences on the performance of the selected models. Among the three studied models, ARMA is the best in terms of filling in the missing monitoring data of air quality. However, it is hard to say which model is better, because the selection of the appropriate model which is based on data properties and the objectives of the analysis, influences to the performance of models. A strange point is that the performance of SVR model in this study is worse than that of ANN and ARMA models. This is different from what reported by several previous studies supposing the need for further studies. Nevertheless, for the prediction of the fluctuation trend of pollutant concentrations, the studied SVR model is better the traditional approaches including LR and Spline interpolation. This study suggested that the machine learning approaches including ARMA, ANN and SVR are potential methods for filling-in the missing values of air quality monitoring data.

Acknowledgements. The authors would like to acknowledge the Center for Environmental Monitoring (CEM), Viet Nam Environment Administration for providing with the data of air quality monitoring stations for this study.

REFERENCES

1. Koutsoyianis D. and Langousis A. - Precipitation, Treaties on water science, ed. Wilderer P. and Uhlenbrook S. Academic Press, Oxford, 2011.
2. Şahin Ü. A., Bayat C., and Uçanc O. N. - Application of cellular neural network (CNN) to the prediction of missing air pollutant data, Atmospheric Research **101** (2011) 314-326.
3. B. H., Raleigh M. S., Fisher A., and Lundquist J. D. - A comparision of methods for filling gaps in hourly near-suface air temperature data, J. Hydrometeorol **14** (3) (2013) 929-945.
4. Alavi N., Warland J. S., and Berg A. A. - Filling gaps in evapotranspiration meaurements for water budget studies: Evaluation of a Kalman filtering approach, Agric. For. Meteorol. **141** (1) (2006) 57-66.
5. Lin K.P., Pai P.F., and Yang S.L. - Forecasting concentrations of air pollutants by logarithm support vector regression with immune algorithms, Applied Mathematics and Computation **217** (2011) 5318-5327.
6. Sánchez A. S., Nieto P. J. G., Fernández P. R., Díaz J. J. d. C., and Iglesias-Rodríguez F. J. - Application of an SVM-based regression model to the air quality study at local scale in the Avilés urban area (Spain), Mathematical and Computer Modelling **54** (2011) 1453-1466.
7. Lu W.-Z. and Wang W.-J. - Potential assessment of the ‘‘support vector machine’’ method in forecasting ambient air pollutant trends, Chemosphere **59** (2005) 693–701.
8. Luna A. S., Paredes M. L. L., Oliveira G. C. G., and Correa S. M. - Prediction of ozone concentration in tropospheric levels using artificial neural networks and support vector machine at Rio de Janeiro, Brazil, Atmospheric Environment **98** (2014) 98-104.
9. Mac Duy Hung, Nghiem Trung Dung, and Dinh Thu Hang. - Application of artificial neural network to fill in the missing monitoring data of air quality, Vietnam Journal of Science and Technology (VAST) **53** (3A) (2015) 199-204.
10. Mac Duy Hung and Nghiem Trung Dung - Application of Echo State Network for the forecast of air quality, Vietnam Journal of Science and Technology (VAST) **54** (1) (2016) 54-63.
11. Mac Duy Hung, Nghiem Trung Dung, and Hoang Xuan Co - Application of Multilayer Perceptron Neural Network for the forecast of tropospheric ozone in Hanoi, Journal of Science and Technology of Technical Universities **111** (2016) 46-51.
12. Neusser K. - Autoregressive moving average models, Time series econometrics. Springer Intenational Publishing, Switzerland, 2016.
13. Ooba M., Hirano T., Mogami J. I., Hirata R., and Fujinuma Y. - Comparisions of gap-filling methods for carbon flux dataset: A combination of a genetic algorithm and artificial neural network, Ecological Modelling **198** (2006) 473-486.
14. Vapnik V. N. - An Overview of Statistical Learning Theory, Proceeding of the IEEE Transactions on Neural Networks **10** (5) (1999) 988-999.
15. Pai P. F. and Hong W. C. - A recurrent support vector regression model in rainfall forecasting, Hydrological Processes **21** (2007) 819-827.